UNDERSTANDING STATISTICS & EXPERIMENTAL DESIGN

Content

1. Basic Probability Theory
2. Signal Detection Theory (SDT)
3. SDT and Statistics I and II
4. Statistics in a nutshell
5. Multiple Testing
6. ANOVA
7. Experimental Design & Statistics
8. Correlations & PCA
9. Meta-Statistics: Basics
10. Meta-Statistics: Too good to be true
11. Meta-Statistics: How big a problem is publication bias?
12. Meta-Statistics: What do we do now?

What do we do now?

Greg Francis – lecture IV

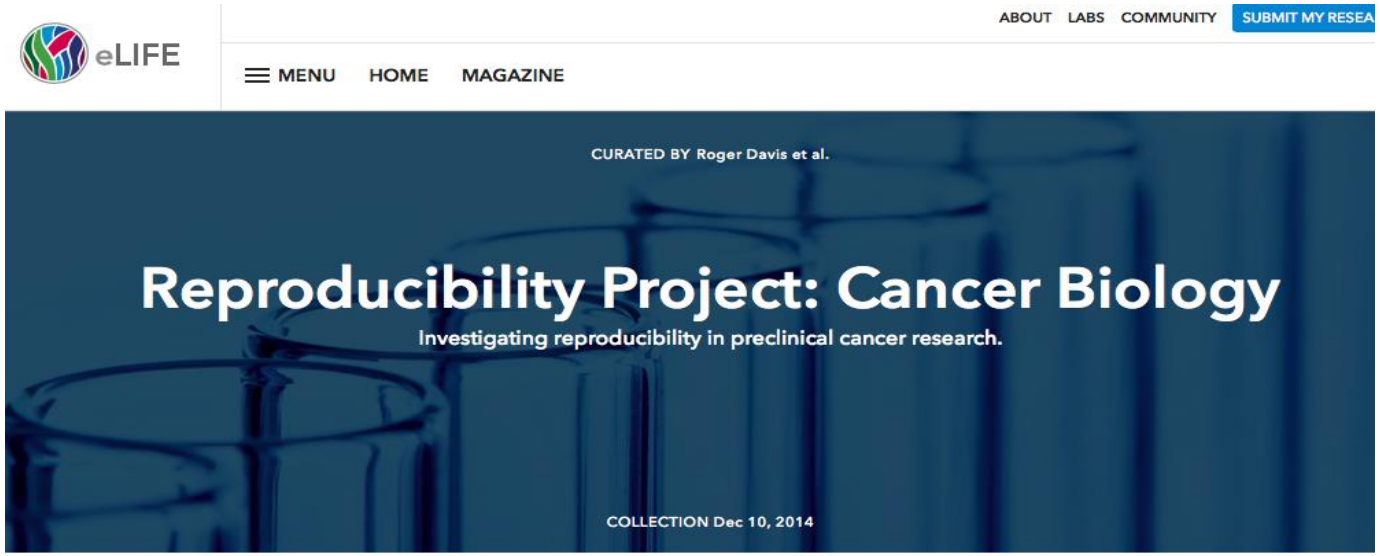
Reproducibility

- Open Science Collaboration (2015) “Estimating the reproducibility of psychological science”. *Science*.
- Conducted replications of 100 studies from three top psychology journals published in 2008
  - *Psychological Science*
  - *Journal of Personality and Social Psychology*
  - *Journal of Experimental Psychology: Learning, Memory, and Cognition*
- Only 36% of replications produced statistically significant effects

	Replications P<.05 in original direction	Percent
Overall	35/97	36
JPSP, social	7/31	23
JEP:LMC, cognitive	13/27	48
PSCI, social	7/24	29
PSCI, cognitive	8/15	53

Reproducibility

- Reproducibility Project: Cancer Biology
- Motived by the report from Amgen that 47 of 53 landmark cancer papers did not replicate
- Conducting independent replications of 50 cancer studies published in *Nature*, *Science*, *Cell* and other high-impact journals
  - Summer 2018: reduced to 18 studies
- As of December 2018:
  - 4 studies reproduced important parts of the original papers
  - 3 studies produced mixed results
  - 2 studies produced negative results
  - 2 studies could not be interpreted



Reproducibility

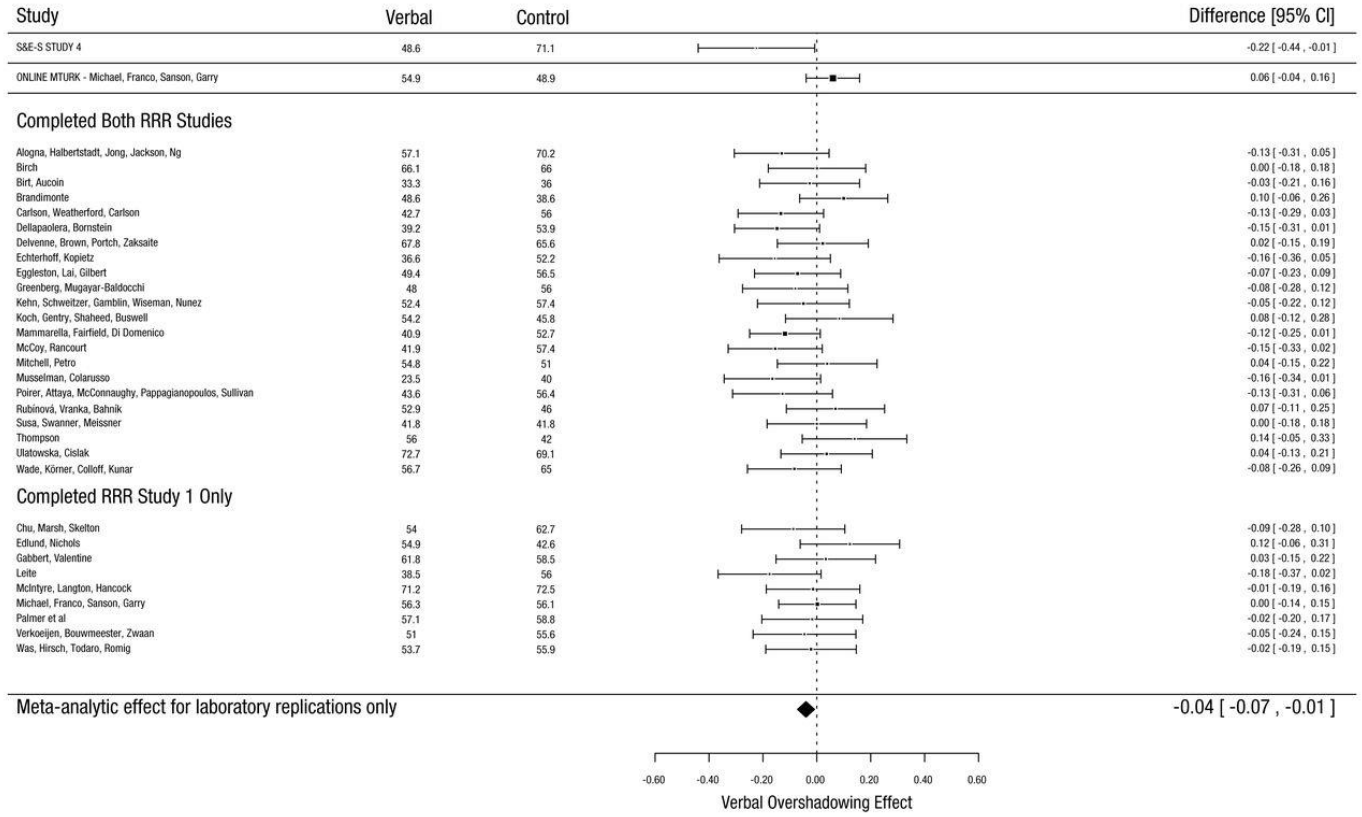
- I find the these reproducibility projects rather strange
  - They are purely empirical investigations of replication
  - The findings are devoid of any theoretical context (for the field in general and for the original studies)
- Still, the low replication rates are disturbing
- There are now calls for new ways of doing science
  1) Do not trust a single study
  2) Trust only replicable findings
  3) Pre-register experiment / analysis designs
  4) Run high powered studies
  5) Lower the *p*-value
  6) Alternative statistics
- I don't think any of these ideas are sufficient, we need to explore fundamental issues

1) Can we trust a single study?

- If it is a well done study, I do not see why not.
  - p=0.01 means the same thing whether the data are from a single study or from a pooling of multiple studies
  - I do not deny value in replication studies, but it is too much to claim that we cannot trust a study by itself
- The value of replication is to test experimental methods
  - Generalization (across samples, equipment, experimenters)
- Such replication tests are very difficult because the comparison of methods requires multiple well done studies with convincing results
  - Science is difficult

2) Should we only trust replicated findings?

- We already saw that too much success (replication) can indicate that the reported findings are biased and should not be trusted
- Moreover, the studies that fail to replicate (e.g., produce non-significant outcomes) still contain some information



2) Should we publish regardless of significance?

- Non-significant experiments contain information about effects
- Publishing both significant and non-significant effects would allow readers to see all the data and draw proper conclusions about effects
- But, readers may not know how to interpret an individual study
  - Can there be a “conclusions” section to an article when it only adds data to a pool?
  - Have to use meta-analysis (how to decide which findings to include?)
- When do you stop running experiments?
  - If the meta-analysis gives $p=0.08$, do you add more experiments until $p<0.05$?
  - That’s just optional stopping at the experiment level
- If you publish all findings for later analysis, why bother with the hypothesis test at all?
- When **would** you do the hypothesis test?

3) Does pre-registration help?

- One suggestion is that experiments should be “pre-registered”
- A description of the experiment and its analysis is posted in some public place before gathering data
  - Limits publication bias
  - Specifies sample size – no optional stopping
  - Specifies experimental measures – cannot “fish” for significant effects
  - Specifies planned analyses / hypotheses – cannot try various tests to get significance
- In as much as it forces researchers to think carefully about their experiment designs, pre-registration is a good idea
- Exploratory work is still fine, one just has to be clear about what it is

3) Does pre-registration help?

- Still, pre-registration strikes me as **unnecessary** or silly
- Extreme case 1 (Unnecessary)
  - Suppose I have a theory that makes a clear prediction (e.g., Cohen's *d*=0.5)
  - I design the experiment to have power of 0.99 (n_1=308, n_2=308)
  - If the experiment produces a significant result, I **can** take that as some support for the theory
  - If the experiment fails to produce a significant result, I **can** take that as evidence *against* the theory
  - But, if I can justify the experiment design with the properties of the theory, then all I have to do to convince others that I had a good experiment is to explain the theory
    - ➢ pre-registration does not improve the experimental design or change the conclusions

3) Does pre-registration help?

- Still, pre-registration strikes me as unnecessary or **silly**

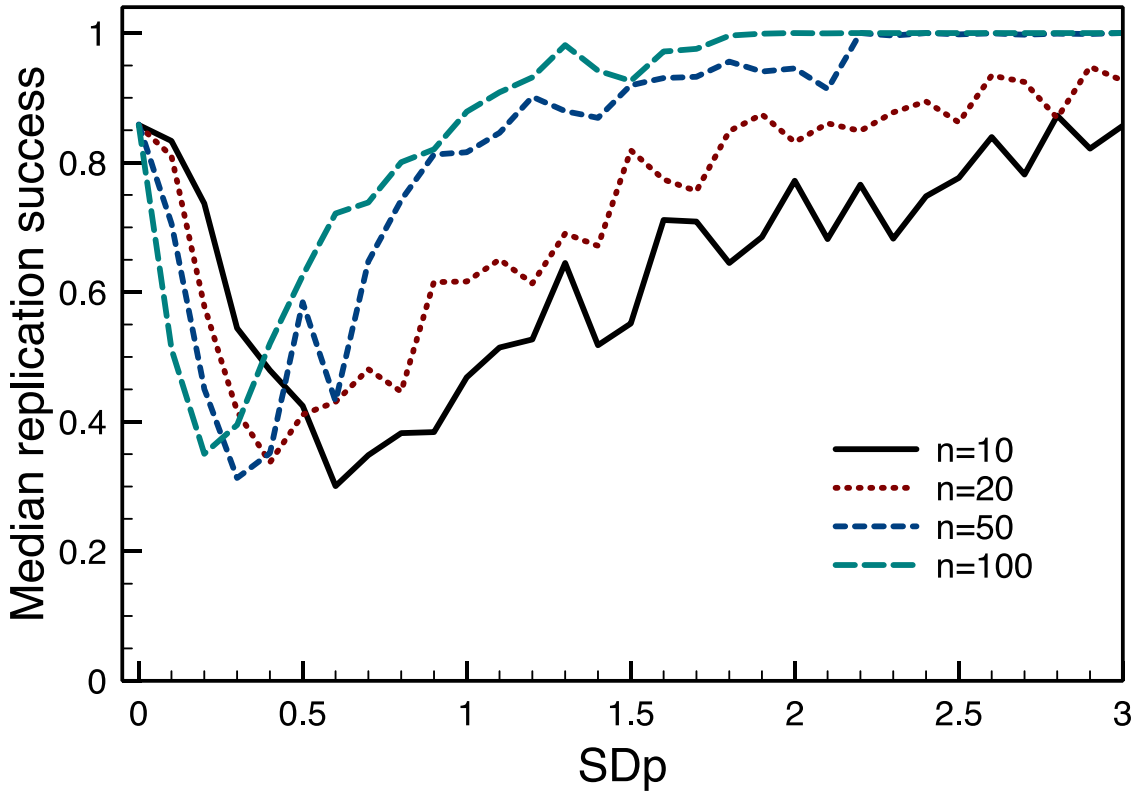
- Extreme case 2 (Silly)
  - Suppose I have no theory that predicts an effect size, but I have a "feeling" that there may be a difference in two conditions for some measure
  - I guess that I need n_1=30 and n_2=30, and I pre-register this design (along with other aspects of the experiment)
  - If the experiment produces a significant result, I **cannot** take that as support for a theory (because there is no theory). The best we can conclude is that my "feeling" may have been right. But that's not a scientific conclusion (no one else can have my "feeling")
  - If the experiment fails to produce a significant result, I **cannot** take that as evidence *against* the "feeling". I never had a proper justification to believe the experiment would work.
  - Pre-registration cannot improve an experiment with unknown design quality

4) Why don't we run high powered studies?

- Suppose you run a 2x2 between subjects design
- You analyze your data to look for main effects and an interaction
  - If there really is no difference in the populations, you have a Type I error rate of 14% for at least one of the tests being significant
- Suppose you just "follow the data" and conclude the presence/absence of an effect as indicated by statistical significance
  - What are the odds that a replication produces the same pattern of results?
- Consider it with a simulation, where you randomly generate 4 population means drawn from a normal distribution with a standard deviation SD_p
  - Then take data samples from normal distributions having those population means
  - Run an ANOVA and draw your conclusions (based on $p<.05$ for main effects and the interaction)
  - Take another sample (of the same size) from the populations and run another ANOVA. See if you get the same conclusions

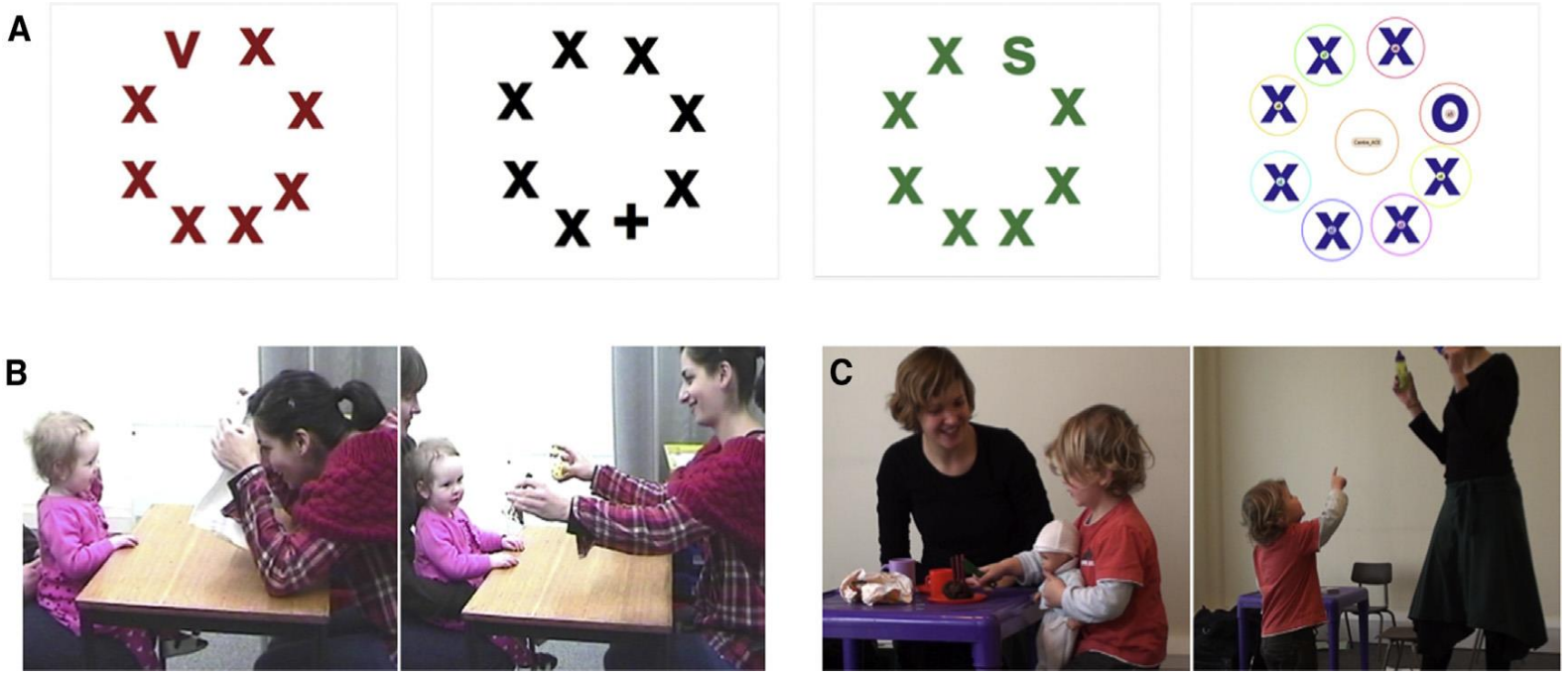
4) Why don't we run high powered studies?

- The replication rate is often low because some of your conclusions are based on findings that are close to the criterion for statistical significance
- Increasing the sample size does not solve this problem, it just shifts it to a different set of conclusions
- In practice, the problem would be even worse
  - A significant interaction would motivate a look for significant contrasts
  - HARKing



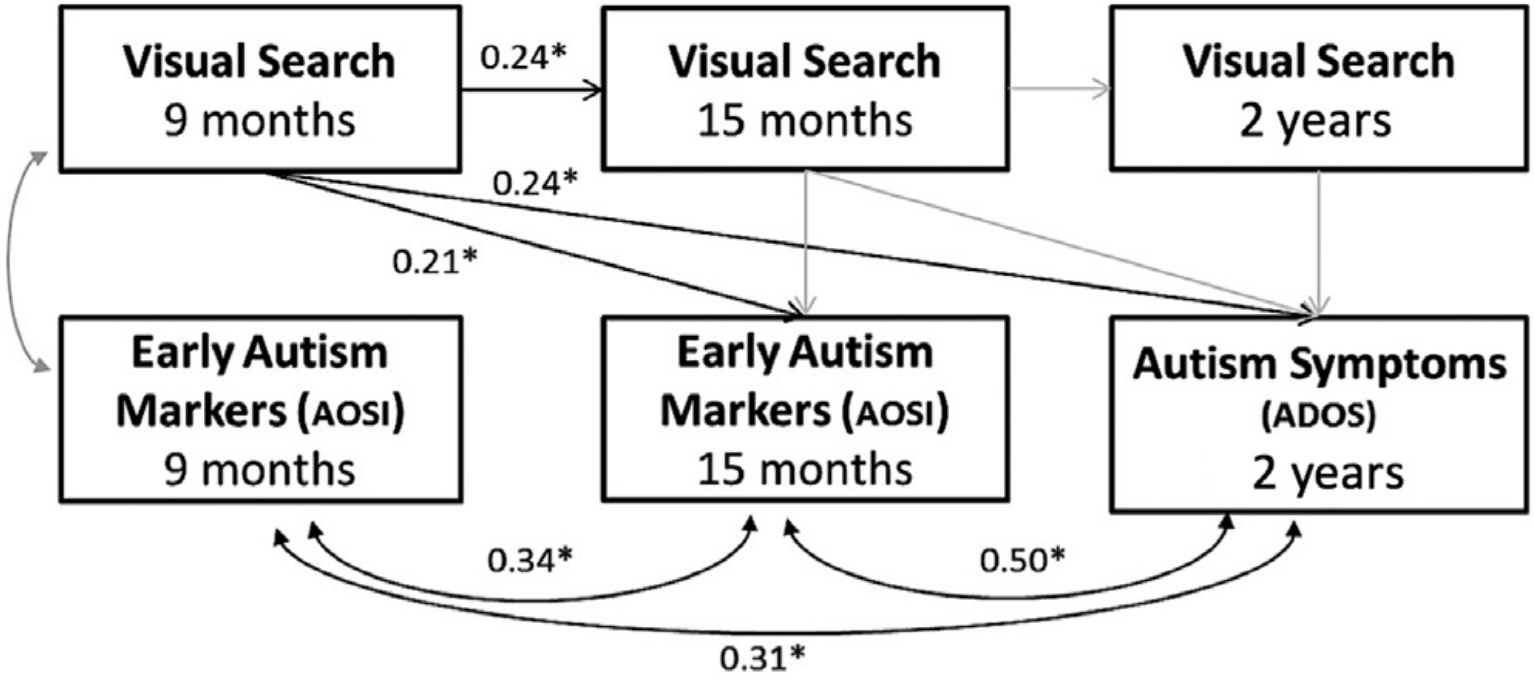
4) Why don’t we run high powered studies?

- Consider a study by Gliga *et al.* (2015, *Current Biology*) on autism and visual search
  - concluded that measurements on a visual search task among infants could predict emerging autism symptoms that appeared months later
  - 82 high-risk for autism infants
  - 27 low-risk control infants



4) Why don’t we run high powered studies?

- Theoretical conclusions were based on significant and non-significant correlations between measurements



4) Why don’t we run high powered studies?

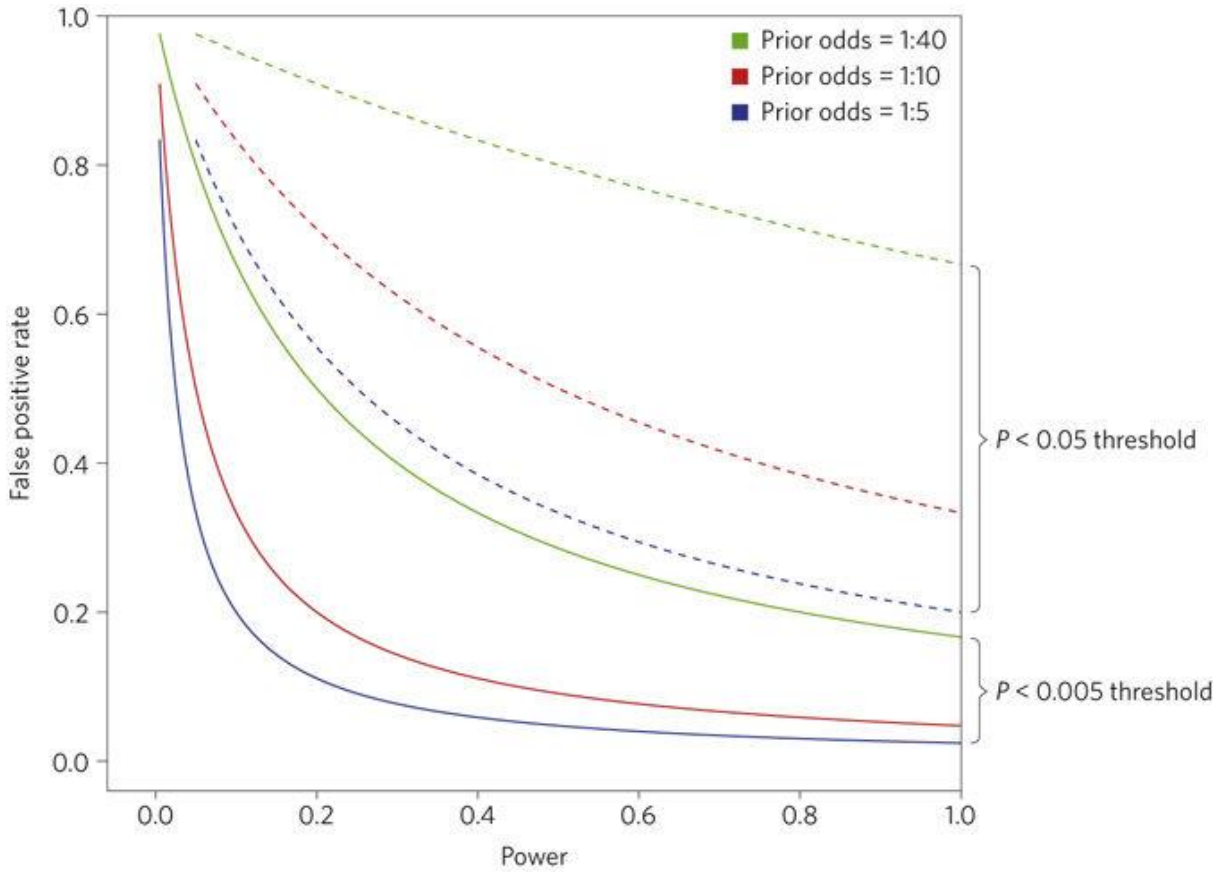
Comparison	r	p	Probability of success
9-month VS and 15-month VS	0.24	0.02	0.676
9-month VS and 9-month AOSI	0.08	0.44*	0.877
9-month VS and 15-month AOSI	0.22	0.03	0.602
9-month VS and 2-year ADOS	0.24	0.02	0.679
15-month VS and 15-month AOSI	NA	>0.05*	0.950
15-month VS and 2-year ADOS	NA	>0.05*	0.949
2-year VS and 2-year ADOS	NA	>0.05*	0.949
9-month AOSI and 15-month AOSI	0.34	0.0007	0.940
9-month AOSI and 2-year ADOS	0.31	0.002	0.888
15-month AOSI and 2-year ADOS	0.50	<0.0001	≈1.00
High-risk infants only			
9-month VS and 15-month AOSI	0.223	0.049	0.454
9-month VS and 2-year ADOS	0.27	0.02	0.524
Partial correlations			
9-month VS and 15-month AOSI; factoring out 9-month AOSI	0.182	0.046	0.546
9-month VS and 2-year ADOS; factoring out 9-month AOSI and 15-month AOSI	0.13	0.13*	0.655
All tests			0.052

4) Why don’t we run high powered studies?

- We do not run high powered studies because we do not want them
- As scientists we want to squeeze the maximum information out of our data set
- That means we will continually break down strong (high powered) effects into more specific effects that (necessarily) have lower power
- Given how we develop our theoretical claims, our experimental findings *should* have low power
- If we want our findings to be reliable, we need a different way of developing theoretical claims

5) smaller p-value?

- Benjamin et al. (2017). Redefine statistical significance. *Nature Human Behaviour*.
- Propose using p<.005 rather than p<.05
- More stringent, fewer false positives
  - Depends on the prior probability of effects being real
- Of course p<.005 and p<.05 are both arbitrary
- Being more stringent means you reduce power for a given effect size and sample size
  - One does not generally promote replication by making success harder to achieve



6) Will alternative statistics help?

- Many people argue that the currently dominant approach for statistical analysis is fundamentally flawed
  - *p*-values, *t*-values

- And that other statistics would be better
  - Standardized effect size (Cohen's *d*, Hedge's *g*)
  - Confidence interval for *d* or *g*
  - JZS Bayes Factor
  - Akaiki Information Criterion (AIC)
  - Bayesian Information Criterion (BIC)

6) Will alternative statistics help?

- In fact, for a 2-sample *t*-test with known sample sizes n_1 and n_2, all of these statistics (and a few others) are *mathematically equivalent* to each other

- Each statistic tells you exactly the same information *about the data set*
  - Signal-to-noise ratio
  - Given one statistic, you can compute all the others
  - http://psych.purdue.edu/~gfrancis/EquivalentStatistics/

- The statistics vary only in how you *interpret* that information

- **You should use the statistic that is appropriate for the inferential interpretation you want to make**

6) Will alternative statistics help?

- No method of statistical inference is appropriate for every situation

- Common statistics are equivalent with regard to the *information* in the data set

- But they can sometimes reach very different conclusions
  - n_1=n_2=250, d=0.183
  - CI_{95} = (0.007, 0.359)
  - p=0.04
  - ΔBIC= - 2.03 (evidence for null)
  - ΔAICc = 2.16 (full model better predicts future data than the null model)
  - JZS Bayes Factor = 0.755 (weak evidence that slightly favors the null model)

Prisoner's Dilemma

- If you behave well (e.g., report null results, do not practice optional stopping), then you produce fewer "amazing" results than someone who behaves badly.
  - Your advisor might be disappointed in your output and will not help you (so much) in your career
- It can take a lot of time to convince established scientists that they are doing some things improperly
- It is certainly true that statistics is not the only important characteristic of science
- I wish I had good advice for how to deal with these issue
- I think some threat of being "caught" is necessary in order to balance the field

Fundamentals

- Statistical significance is not the long-term goal of science

- Often want to put ourselves in a situation where statistics hardly matter at all (look at physics)
  - Then replication helps identify methodological (rather than statistical) issues

- DARPA recently put out a call for methods to identify “robust effects” in the social sciences
  - Wisdom of the crowds
  - Replication
  - Researcher reputation
  - Pattern of citations
- How do we know something is “robust”?

Not Just Replication

- Consider superconductivity
- Discovered in 1911, plays an important role in fMRI
- How do we know superconductivity works the same in Lausa
  - Mountains?
  - Lake versus river?
  - Brick buildings?
  - French versus English?
  - 7T versus 3T?
- It's not just that superconductivity worked before!
  - Every new environment is different

Mechanisms

- There is a theory about superconductivity that describes mechanisms that produce it
  - Meissner effect (1930s)
  - Cooper pairs in quantum mechanisms (1950s)
- This theory predicts when superconductivity works and when it does not
  - That's how engineering works
- It's not perfect
  - High-temperature superconductivity remains unexplained
  - That's where science is being done
- We know/believe fMRI will work in both Lausanne and West Lafayette because we understand the mechanisms that determine when superconductivity will happen

Getting To Mechanisms

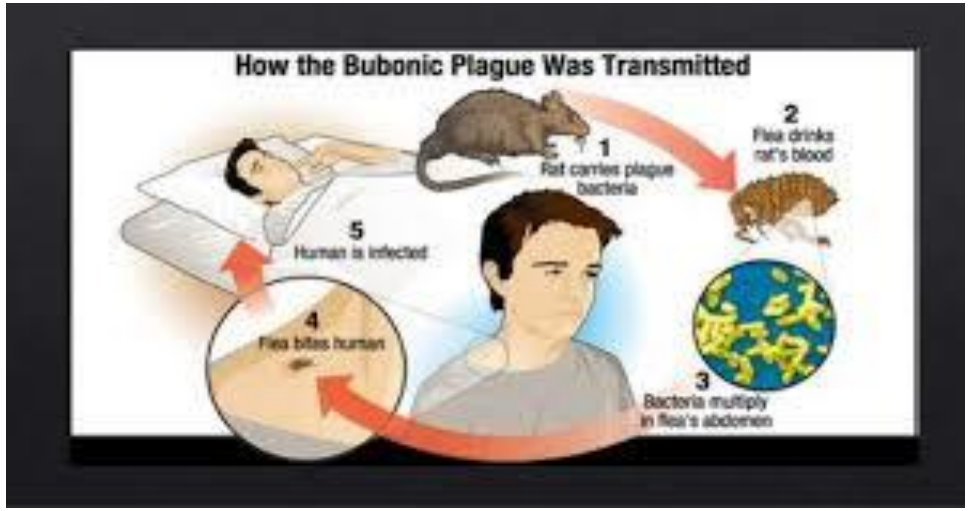
- If we want to have successful/robust science, our long term goal is identification of mechanisms

- We might not get there in our lifetime
  - Exploratory work
  - Confirmatory work
  - Proposing theories
  - Testing theories

- We are not going to have replicable/reproducible science until we can specify and justify mechanisms
  - It might not look the same as in physics

Plague

- Paul-Louis Simond (1898) discovered that the plague was transmitted by fleas on rats
- Once a mechanism is identified, it suggests what to do
- To reduce occurrence of the plague, reduce the number of rats and contact with rats
  - Kill rats
  - Keep dogs and cats
  - Seal food containers
  - Set rat traps
  - Avoid rats
  - Don’t bother with quarantining the family of an infected person



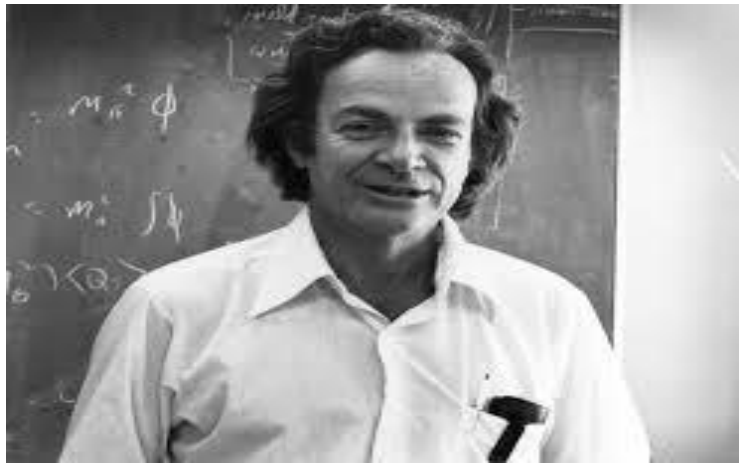
- Not precise predictions about the magnitude of the benefit (but they should all work to some extent)

Mechanisms In Life Sciences

- Psychology faces challenges because there are very few proposed mechanisms
  - Even when something seems to be a strong effect, we cannot judge when it will apply and when it will not
- Neuroscience and medicine have some hope because scientists naturally seek out mechanisms based on biology
  - But there are other problems with sample sizes and costs of investigations
- Keep in mind that the long-term goal is to identify mechanisms, and plan studies and analyses accordingly

Conclusions

- There are methods that identify improper or invalid scientific analyses
- *"Only when the tide goes out do you discover who has been swimming naked." Warren Buffet*
- Science requires a level of care that goes well beyond most professions
- Nobel prize winning physicist Richard Feynman described it this way:
  - "**The first principle is that you must not fool yourself--and you are the easiest person to fool.**"
- Be careful. Be honest.



Take home messages

Take Home Messages

1. Many suggestions, such as preregistration, to improve statistical practice do not address the fundamental problems.
2. Good science involves more than just statistics.

END Class 12